DOCUMENT RESUME

ED 452 207 TM 032 492

AUTHOR Capraro, Mary Margaret; Capraro, Robert M.; Henson, Robin K.
TITLE Measurement Error of Scores on the Mathematics Anxiety Rating Scale across Studies.
PUB DATE 2001-02-00
NOTE 37p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (New Orleans, LA, February 1-3, 2001).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Adults; Mathematics Anxiety; Meta Analysis; *Rating Scales; *Reliability; *Scores
IDENTIFIERS *Mathematics Anxiety Rating Scale

ABSTRACT

The Mathematics Anxiety Rating Scale (MARS) (F. Richardson and R. Suinn, 1972) was submitted to a reliability generalization analysis to characterize the variability of measurement error in MARS scores across administrations and to identify possible study characteristics that are predictive of reliability variation. The meta-analysis was performed with 67 studies that met study criteria. In general, the MARS and its variants yielded scores with strong internal consistency and test-retest reliability estimates, although variation was observed. Adult samples were related to lower score reliability compared to other age groupings. Inclusion of total score standard deviation in the regression models resulted in roughly 25% increases in R squared effects. (Contains 3 tables and 44 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

TM

ED 452 207

Running head: MATHEMATICS ANXIETY

Measurement Error of Scores on the Mathematics Anxiety Rating Scale Across Studies

Mary Margaret Capraro and Robert M. Capraro

Texas A&M University

Robin K. Henson

University of North Texas

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

M. Capraro

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Paper presented at the annual meeting of the Southwest Educational Research Association, February 1-3, 2001, New Orleans. Correspondence to the first and second authors should be sent to mmcapraro@coe.tamu.edu and rcapraro@coe.tamu.edu, respectively. Correspondence to the third author should be sent to rhenson@tac.coe.unt.edu.

TM032492

BEST COPY AVAILABLE

Abstract

The Mathematics Anxiety Rating Scale (MARS) was submitted to a reliability generalization analysis to characterize the variability of measurement error in MARS scores across administrations and identify possible study characteristics that are predictive of reliability variation. In general, the MARS and its variants yielded scores with strong internal consistency and test-retest reliability estimates, although variation was observed. Adult samples were related to lower score reliability compared to other age groupings. Inclusion of total score standard deviation in the regression models resulted in roughly 25% increases in $\underline{R}^2$ effects.

Measurement Error of Scores on the Mathematics Anxiety Rating Scale Across Studies

Regarding measurement error, it is important to emphasize that scores, not tests, are either reliable or unreliable (Thompson, 1994; Vacha-Haase, 1998). As correctly noted by Gronlund and Linn (1990), "Reliability refers to the results obtained with an evaluation instrument and not to the instrument itself. Thus it is more appropriate to speak of the reliability of 'test scores' or the 'measurement' than of the 'test' or the 'instrument'" (p. 78, emphasis in original). Many researchers, however, unfortunately refer to the "reliability of the test." This phraseology may lead many to incorrectly assume that reliability inures to tests rather than scores, and can result in researchers often failing to examine score reliability for their data. These points, and others, have been vociferously discussed. As examples, Thompson and Vacha-Haase (2000) presented a case for characterizing reliability in terms of scores, not tests. Sawilowsky (2000) presented a contrary view.

The argument that reliability is a function of scores and not the test itself is not mere semantics. Indeed, there are important research implications of the view that score reliability may vary across administrations of a measure. For example, poor score reliability can attenuate observed effect

sizes. As Reinhardt (1996) observed:

> Reliability is critical in detecting effects in substantive research. For example, if a dependent variable is measured such that the scores are perfectly unreliable, the effect size in the study will unavoidably be zero, and the results will not be statistically significant at any sample size, including an incredibly large one. (p. 3)

Accordingly, poor reliability can reduce statistical power (Onwuegbuzie & Daniel, 2000) and potentially lead to inappropriate conclusions concerning substantive research findings (Thompson, 1994).

Because reliability may fluctuate, researchers should always examine the reliability of their data in hand and report it. The APA Task Force on Statistical Inference agreed, and in a recent report noted:

> It is important to remember that a test is not reliable or unreliable. Reliability is a property of the scores on a test for a particular population of examinees. . . Thus, authors should provide reliability coefficients of the scores for the data being analyzed even when the focus of their research is not psychometric. (Wilkinson & APA Task Force on Statistical Inference, 1999, p. 596)

Furthermore, Henson, Kogan, and Vacha-Haase (in press) emphasized that,

> It is insufficient to assume that a test will yield reliable scores solely because reliable scores have been obtained in the past. An even more egregious error is to assume a test will yield reliable scores when reliability has been marginal in the past. . .

Because reliability is a function of scores, any sample characteristic that can affect scores can impact reliability. For example, Thompson (1994) observed that "The same measure, when administered to more heterogeneous or more homogeneous sets of subjects, will yield scores with differing reliability" (p. 839). If we assume that a sample is heterogeneous as regards the trait of interest, then the subjects will likely score differently from each other, resulting in increased total score variance (at least to the degree of heterogeneity assumed). Classical test theory estimates (e.g., coefficient alpha, test-retest) assume that increased total variance indicates a more reliable (accurate) measure for each person because the likelihood decreases that a person's rank ordering in the distribution would change if measured again.

Because heterogeneous samples will tend to yield larger total variance, tests given to such samples will tend to yield higher reliability estimates. This clearly is a function of the characteristics of the sample and not the test <u>per se</u>. As such, Reinhardt (1996) explained that "both the characteristics of the

person sample selected and the characteristics of the test item can affect coefficient alpha" (p. 6). Furthermore, Dawis (1987) emphasized that "reliability is a function of sample as well as of instrument, [reliability] should be evaluated on a sample from the intended target population - an obvious but sometimes overlooked point" (p. 486). Score reliability, then, may vary depending on the characteristics of the sample from which the scores are obtained, including differential impact from homogeneous versus heterogeneous sample compositions.

Estimating Fluctuation of Reliability Estimates

Because score reliability can (and will) vary from study to study, Vacha-Haase (1998) presented reliability generalization (RG) as a methodology for examining measurement error variance across studies. Based on validity generalization methods (Hunter & Schmidt, 1990; Schmidt & Hunter, 1977), RG studies can provide information regarding: a) the variability of score reliability estimates across administrations of a measure, and b) the substantive study characteristics that may affect those reliability estimates.

Essentially, any measure that has some frequency of use in the literature can be submitted to a RG analysis. However, because RG often uses reliability estimates as the central dependent variable, only those studies reporting reliability can eventually find their way into the analysis. As Thompson and

Vacha-Haase (2000) noted, ". . . the RG chef can only work with the ingredients provided by the literature" (p. 184). Of course, RG has not been characterized as a monolithic method, and can involve a variety of information that may be used to describe psychometric properties of scores (e.g., coefficient alpha, standard error of measurement, etc.). As more authors report such information, there may exist more "fodder for reliability generalization analyses focusing upon the differential influences of various sources of measurement error" (Vacha-Haase, 1998, p. 14).

Despite the recency of RG methodology, several RG studies are now present in the literature. As RG studies continue to be conducted, and published, the field will hopefully develop cumulative knowledge of: a) the degree score reliability varies for instruments, and b) whether study characteristics can consistently predict measurement error for a test or perhaps even across tests or constructs. Examples of RG studies include examinations of the Bem Sex Role Inventory (Vacha-Haase, 1998), Beck Depression Inventory (Yin & Fan, 2000), "Big Five Factors" of personality across various tests (Viswesvaran & Ones, 2000), NEO-Five Factor Inventory (Caruso, 2000), White Racial Identity Attitude Scale (Helms, 1999), and Teacher Efficacy Scale (Henson et al., in press).

Purpose

The purpose of the present study was to conduct a meta-analytic RG study on the Mathematics Anxiety Rating Scale (MARS; Richardson & Suinn, 1972), the leading instrument used to assess self-reported anxiety toward mathematical content and performance. Reliability estimates (coefficient alpha and test-retest) were examined to characterize the typical reliability for multiple administrations of the MARS. Study characteristics (e.g., sample size, gender of participants, test length) were investigated as possible predictors of score reliability variation.

Mathematics Anxiety Rating Scale

The MARS (Richardson & Suinn, 1972), originally a 98-item inventory, was constructed to provide a unidimensional measure of anxiety associated with the manipulation of numbers and the use of mathematical concepts. The instrument contains short descriptions of real-world and academic situations that may stimulate mathematics anxiety. Participants record their responses on a five-point Likert scale ranging from one (none at all) to five (very much). On the original version, the item scores are summed to give a total range of 98 to 490, with higher scores reflecting higher mathematics anxiety. It should be noted that some of the initial tests were inadvertently

published with only 94 items, thus test length may vary even for the original version of the MARS.

Although the MARS is the most commonly used measure of mathematics anxiety, related instruments include the Fennema-Sherman Mathematics Anxiety Survey (Fennema & Sherman, 1976), Dreger and Aiken's (1957) Numerical Anxiety Scale, and the Mathematics Anxiety Questionnaire (Wigfield & Meece, 1988). The MARS has become the most popular instrument used in the area due to its extensive data on the reliability and validity of scores from the scale (Plake & Parker, 1982). In order to broaden applicability across age groups, the MARS has been periodically revised, including the MARS-E (Suinn, 1988) and MARS-A (Suinn & Edwards, 1982) for elementary and adolescent students, respectively. The popularity of the test has encouraged other researchers to develop revised forms of the original MARS. Some examples of attempts to develop shortened versions include a 24-item test by Plake and Parker (1982) and a 25-item test by Alexander and Martray (1989).

MARS Score Reliability

Reliability of scores on the MARS is reported by some to be relatively high (Alexander & Martray, 1989). The MARS normative data (Richardson & Suinn, 1972; Suinn, Edie, Nicoletti, & Spinelli, 1972) indicated a 2-week test-retest reliability of .78, a seven-week test-retest with a second sample of .85, and

an internal consistency (alpha) on the second sample of .97. Data provided in the MARS Informational Brief (R.M. Suinn, personal communication, March 27, 2000) indicated a two-week test-retest reliability of .86 for women, .95 for men, and .87 for the total sample. Coefficient alphas were reported as .97 for women, .99 for men, and .96 for the total sample.

MARS Score Validity

Validity of scores for the original version was established in two ways. First, from a construct validity perspective, high mathematics anxiety should be associated with lower performance on mathematics tests. Richardson and Suinn (1972) claimed evidence of construct validity based on a study of 30 students enrolled in an advanced undergraduate psychology class. Roughly equally divided between males and females, the students completed the MARS and were then administered the Differential Aptitude Test (DAT; a commonly used test to assess mathematics ability). The correlation between MARS and DAT scores was -.64, indicating that greater anxiety was associated with poor performance on the DAT.

Second, clinical subjects treated for mathematics anxiety in three separate studies showed scores above that of the normal standardization MARS samples. Following treatment for mathematics anxiety, the treated subjects' scores showed decreases as compared with untreated subjects. Assuming that the

treatment program did in fact reduce the level of mathematics anxiety, the change in MARS' scores may be viewed as providing construct validity evidence.

Although these studies report adequate reliability and validity of scores from the MARS, as noted above reliability (and validity) can fluctuate on subsequent samples. The present study examines the measurement error fluctuation of MARS scores across published studies.

Method

Article Selection

A search for articles using the MARS was conducted in the ERIC and PsycLit databases using the keyword "mars" from 1970 to June 2000. A total of 226 articles were identified from the ERIC database and 118 from PsycLit. Of this total (344), many articles were false hits and two were unable to be obtained, leaving 83 articles that actually used the MARS (43-ERIC, 40-PsyLit). After eliminating duplicate articles (and possible conference presentations in ERIC) between the databases (16), 67 articles remained in the sample. These articles were then coded for multiple criteria including whether they reported a reliability estimate. Of these 67, only 17 (25%) reported at least one reliability estimate for the data in hand. However, some of these articles reported more than one estimate as part of separate samples or sample subgroups. Each of these estimates

was treated as a separate case in the data analysis, yielding 35 total reliability coefficients. Of these 35, 7 were test-retest and 28 were coefficient alpha estimates.

Coding of Study Characteristics

The 67 articles using the MARS (17 of which actually reported reliability) were read and coded on multiple criteria intended to capture study characteristics that may impact score reliability. Specifically, many characteristics were framed such that they may describe features that would suggest sample homogeneity. These features were examined because classical test theory reliability estimates are impacted by the total test score variance, and it has been shown that as subjects score differently (i.e., as samples are more heterogeneous) reliability tends to increase (cf. Reinhardt, 1996; Thompson, 1999; Henson, 2000). As Henson et al. (in press) explained:

> In terms of classical measurement theory (holding the number of items on the test and the sum of item variances constant), increased variability of total scores suggests that we can more reliably order people on the trait of interest, and thus more accurately measure them. This assumption is made explicit in the test-retest reliability case, when consistent ordering of people across time on the trait of interest is critical in obtaining high reliability estimates.

Although multiple study characteristics were coded, the small percentage of studies actually reporting reliability coefficients limited the number of variables that could be used and the types of analyses conducted. When coefficient alpha was reported, information on several predictors was either not given or insufficiently reported. After listwise deletion for missing data, several predictors were omitted from further analyses to maintain an adequate sample size. Many of the remaining coded variables selected for analysis had particular potential for capturing differences in sample homogeneity. The coded variables were:

1. Number of items on the test.
2. Number entries on the Likert scale: 4 = four point scale, 5 = five point scale.
3. Sample size for the reliability coefficient reported.
4. Age of sample. Five dummy coded variables were created that contrasted: all children, all adolescents, all college age, all adults, and mixed ages (all coded 1) versus all other groups (0). These five dummy vectors were treated as separate variables in the analyses because of the typical application of the MARS, in which the test is often administered to homogenous age groups to assess anxiety levels.
5. Gender homogeneity: Coded as proportion of the number of persons in the majority gender to total sample size. As such,

this variable ranges from .50 to 1.00. This proportion measures gender homogeneity, regardless of whether that homogeneity was due to females or males.

6. Standard deviation of total scores: All standard deviations were given at the sum of total scores level.

7. Ethnicity: 1 = mixed, 0 = homogeneous groups, including all White, all African-American, all Hispanic, all Asian, all Native American, all International.

8. Type of reliability coefficient: 1 = alpha, 2 = test-retest.

Data Analyses

The typical magnitude and variability of reliability estimates was evaluated with descriptive statistics. A series of four multiple regression analyses were conducted to evaluate whether the predictors could account for variation in the reliability estimates. The first regression model included the number of items, Likert scale, sample size, the five dummy coded age variables, and type of reliability estimate as the predictors. Because test-retest estimates tend to be lower than internal consistency reliabilities, the second model included all of the above predictors except for type of reliability and only used the 28 internal consistency estimates (alpha) as the dependent variable. The third model used the same predictors as model 1 but added the total score standard deviation to evaluate the additional effect of total score variance on all reliability

estimates. The fourth model also included the standard deviation but, like model 2, omitted the type of reliability predictor and only used alphas as the dependent variable.

The total score standard deviation was not included in the first two models because some cases did not report this basic information and listwise deletion would have limited the sample size. Inclusion of total score standard deviation was relegated to subsequent models (3 and 4) with lower sample sizes. The focus on alpha only in models 2 and 4 mirrors the approach used in Yin and Fan's (2000) RG on the Beck Depression Inventory, in which type of reliability was found to be a strong predictor of reliability variance (with test-retest estimates generally lower than internal consistency). Unlike the Yin and Fan study, there were not enough test-retest coefficients in the present study (n = 7) to warrant regression with test-retest reliability only. Listwise deletion was used for all multiple regression analyses.

In addition, bivariate correlations were conducted between the reliabilities (both types combined and then alpha only) and the gender homogeneity and ethnicity variables. These two predictors were not included in the multiple regressions due to excessive missing data, which after listwise deletion, would have excessively lowered the number of cases useable in the regression. Their bivariate correlations with the reliabilities are reported separately.

Results

Overall, the MARS tended to yield scores with high reliability (see Table 1). When the coefficients were examined by reliability type, coefficient alpha yielded higher estimates than test-retest estimates. This finding highlights the well-known difficulty of obtaining accurate scores across time in the test-retest case. The Table 1 results point to the ability of the MARS to yield generally acceptable, even high, reliability estimates. However, it is also apparent that even when most estimates are elevated across studies, there still exists measurement error fluctuation and the possibility of lower estimates in a given sample, as evidence by the .550 internal consistency coefficient. The sample for which this estimate was derived (Wilson, 1997) consisted of psychology graduate students enrolled in a testing and individual analysis course - arguably a relatively homogenous group as regards mathematics anxiety.

INSERT TABLE 1 ABOUT HERE

Table 2 presents descriptive statistics for the coded predictors. Because the predictors change across the models and the sample sizes vary due to listwise deletion, descriptives are given for all four models used in the subsequent regression analyses. Examination of Table 2 indicates that all predictors appeared to have reasonable variance except the mixed age group,

whose means were near zero across all four models, indicating that there were few studies actually reporting reliability coefficients for mixed age groups. This finding is consistent with the typical application of the MARS, where specific age groups are generally targeted for evaluation of mathematics anxiety. It is also worth noting that the number of items used in the MARS varied considerably across studies, suggesting that researchers have taken liberty at deleting, or at least ignoring, items from the original 98-item version (Richardson & Suinn, 1972). Furthermore, it is apparent that the majority of the studies used a 5-point Likert scale. In fact, only a children's version used by Chiu and Henry (1990) used a 4-point scale.

INSERT TABLE 2 ABOUT HERE

Table 3 presents results from the regression analyses. The college age predictor was deleted from the analysis in models 1, 2, and 3 due to tolerance limits. Conversely, the children age predictor was deleted from model 4 due to tolerance limits. Looking at Table 2, we find that model 1 yielded a 40.4% effect. The beta weights and structure coefficients indicated a substantial negative relationship between the adult age predictor and the reliability estimates, suggesting that the homogeneous adult samples tended to yield lower reliability

estimates when compared to the other age groups. Furthermore, this pattern was generally consistent across all age groups, although most of the relationships were weak. As expected and consistent with classical test theory, the type of reliability coefficient (alpha versus test-retest) was also a strong predictor of the dependent variable (cf. Yin & Fan, 2000). Test-retest coefficients tended to be lower than the internal consistency estimates.

INSERT TABLE 3 ABOUT HERE

Model 2 examined internal consistency estimates only as the dependent variable. Because model 1 showed a substantial effect based on type of reliability, model 2 was expected to have a lower overall R^2. The model 2 effect was lower (6.2% less than model 1) but remained substantial at 33.2%. Again, the adult age predictor was a primary contributor to the explained variance in the alpha estimates. The structure coefficient for the number of items on the test indicated that this predictor also had a moderate positive relationship to the predicted synthetic variable, a finding consistent with classical test theory.

To examine the impact of sample variance (a proxy estimate of individual differences or sample heterogeneity on the trait of interest), the third model added the total score standard deviation predictor. The large effect observed (64.4%)

represented a sizeable increase in the predicted variance (24.0%) over the model 1 effect. Again, reliability type was the dominant predictor but the betas and structure coefficients also suggested contributions by the number of items on the test and the total score standard deviation. Oddly, however, there was a negative relationship between the number of MARS items and reliability, indicating that reliability estimates tended to decrease as test length increased. Closer examination of the data revealed that the longer tests were associated with the three test-retest estimates. Because test-retest estimates are generally lower than internal consistency estimates, the negative relationship for test length in model 3 speaks more to differences between reliability estimates than the impact of test length on MARS score accuracy in general. When only alpha was examined in model 4, the relationship returned positive.

Finally, prediction of alpha only in model 4 again yielded a large effect (58.2%) with minimal reduction from model 3 (6.2% less). Furthermore, the model 4 effect represented a 25.0% increase over model 2, which also predicted internal consistency estimates only but without total score standard deviation in the model. The adult age predictor was again important along with the college group, number of test items, and standard deviation.

Bivariate correlations were conducted between the gender homogeneity and ethnicity predictors and reliability estimates.

Gender homogeneity was essentially unrelated to score reliability when both alpha and test-retest were considered (r = .141, n = 19) and when alpha only was predicted (r = .100, n = 15). Ethnic homogeneity, however, was negatively related to internal consistency estimates (no test-retest coefficients were available after pairwise deletion) with r = -.643 (n = 11). Because ethnicity was coded as 1 for mixed and 0 for all homogeneous groups, the correlation indicated that alpha tended to decrease with samples of heterogeneous ethnicity. This finding is not consistent with the expectation that heterogeneous samples would yield higher classical test theory reliability estimates. It does indicate that, like gender, the reliability of MARS scores apparently is not negatively impacted by ethnic homogeneity.

Discussion

The articles examined in the present investigation demonstrated that the MARS (and its multiple test length versions) tends to yield scores with strong reliability across administrations. However, like all measures, MARS scores are dependent on sample characteristics, which translates to fluctuating reliability estimates to some degree. For example, despite overall strong coefficient alpha estimates, one study reported a marginal alpha of .550 for MARS scores. This variability in score reliability demonstrates that the most

relevant reliability estimate for one's sample data is the one computed on one's sample data. Therefore, researchers ought to both report and interpret their obtained reliabilities in practically all studies (cf. Henson et al., in press; Thompson, 1994; Thompson & Vacha-Haase, 2000; Vacha-Haase, 1998; Wilkinson & APA Task Force on Statistical Inference, 1999). As Henson et al. (in press) observed, ". . . the best evidence of adequate score reliability for one's own data is to actually compute it - a process that takes at least a minute with modern computing capabilities!"

Regarding study characteristics, there was a consistent pattern for the adult age group variable to be negatively related to reported reliability across all regression models, indicating that the adult samples tended to yield less reliable scores. Most other age based variables were either unrelated or slightly negatively related to reliability. It is possible that adults tend to score more similarly on mathematics anxiety than other age groups, resulting in lower score reliability. As expected, test length was positively related to the dependent variable except in model 3. The model 3 finding, however, was an artifact based on data features discussed above. The Likert scale used was unrelated to reliability. Sample size was also not predictive of reliability variation, with the exception of model 4 where a small negative relationship was observed. This

finding is consistent with Viswesvaran and Ones' (2000) RG on the "Big Five Factors" of personality which indicated no relationship between sample size and reliability. Henson et al. (in press) noted inconsistent levels of prediction by sample size. Of course, various measures may be differently impacted sample characteristics. As RG studies continue, however, it is expected that sample size will be generally not predictive of reliability variation, at least for samples of moderate size.

What is most notable in the present results is the impact of adding total score standard deviation to the overall effect sizes across the regression models. Models with standard deviation included increased $\underline{R}^2$ by 24.0% and 25.0% over the respective models without standard deviation used as a predictor. This finding highlights the potential impact of total score variance on reliability estimates. Classical estimates such as coefficient alpha hinge on the total score variance as an indication of the degree subjects have been reliably measured. While total score variance is not the only data feature taken into account by coefficient alpha, it is clearly a central element in the outcome of the formula (cf. Henson, 2000; Reinhardt, 1996; Thompson, 1999).

An important point here concerns those studies that only reference reliabilities reported in prior studies or the test manual as somehow being relevant for their own data. This

practice, called "reliability induction" by Vacha-Haase, Kogan, & Thompson (2000) due to researchers' attempts to induct a specific reliability estimate to a broader context of studies, may be legitimate only if the inducted sample is similar to the sample under investigation in terms of "composition and variability" (Crocker & Algina, 1986, p. 144). Unfortunately, Vacha-Haase et al. (2000) observed dramatically different sample compositions for published studies as compared to the normative groups on the Bem Sex Role Inventory. These findings are consistent with Dawis' (1987) observation that reliability "should be evaluated on a sample from the intended target population - an obvious but sometimes overlooked point" (p. 486).

In sum, measurement error in MARS scores appears to increase in adult samples and perhaps in other homogeneous age groups. This finding is particularly relevant for the MARS as this test is typically used with specific ages in the assessment of mathematics anxiety. Nevertheless, the MARS demonstrated generally strong score reliability across the administrations studied here. Of course, the many articles that failed to report appropriate reliability may have otherwise impacted the current findings had they been included. The present findings are, therefore, limited by a potential reporting bias toward high reliability estimates and by the relatively small sample sizes

available for analysis due to lack of reporting. Consistent with classical formulations, it is clear that total score variance impacts reliability when added to the predictive models. Based on these results and an understanding of what data features impact reliability estimates, researchers employing the MARS are encouraged to: a) explicitly compare their sample composition and variability to that of the normative sample if referencing the normative sample reliability estimates; or better yet, b) calculate, report, and interpret the reliability of the scores obtained from the sample under investigation.

References

Articles used in the meta-analysis are marked with an asterisk.

*Alexander, L., & Martray, C. (1989). The development of an abbreviated version of the Mathematics Anxiety Scale. Measurement and Evaluation in Counseling and Development, 22, 143-150.

*Bessant, K. (1995). Factors associated with types of mathematics anxiety in college students. Journal for Research in Mathematics Education, 26, 327-345.

*Bush, W. (1989). Mathematics anxiety in upper elementary school teachers. School Science and Mathematics, 89, 499-509.

Caruso, J. C. (2000). Reliability generalization of the NEO personality scales. Educational and Psychological Measurement, 60, 236-254.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston.

*Chiu, L. & Henry, L. (1990). The development and validation of the Mathematics Anxiety Rating Scale for Children. Measurement and Evaluation in Counseling and Development, 23, 121-127.

*D'ailly, H., & Bergering, A. (1992). Mathematics anxiety and mathematics avoidance behavior: A validation study of two

MARS factor-derived subscales. Educational and Psychological Measurement, 52, 369-377.

Dawis, R. V. (1987). Scale construction. Journal of Counseling Psychology, 34, 481-489.

*Dew, K., & Galassi, J. (1983). Mathematics anxiety: Some basic issues. Journal of Counseling Psychology, 30, 443-446.

Dreger, R., & Aiken, L. (1957). The identification of number anxiety in a college population. Journal of Educational Psychology, 48, 344-351.

Fennema, E., & Sherman, J. (1976). Fennema-Sherman Mathematics Attitude Scales: Instruments designed to measure attitudes toward learning mathematics by females and males. JSAS Catalog of Selected Documents in Psychology, 6, 31.

Gronlund, N. E., & Linn, R. L. (1990). Measurement and evaluation in teaching (6th ed.). New York: Macmillan.

Helms, J.E. (1999). Another meta-analysis of the White Racial Identity Attitude Scale's Cronbach alphas: Implications for validity. Measurement and Evaluation in Counseling and Development, 32, 122-137.

Henson, R. K. (2000, November). A primer on coefficient alpha. Paper presented at the annual meeting of the Mid-South Educational Research Association, Bowling Green, KY. (ERIC Document Reproduction Service No. forthcoming)

Henson, R. K., Kogan, L. R., & Vacha-Haase, T. (in press).

A reliability generalization study of the Teacher Efficacy Scale and related instruments. Educational and Psychological Measurement.

Hunter, J. E., & Schmidt, F. L. (1990). Methods of meta-analysis. Newbury Park, CA: Sage.

*McAuliffe, E., & Trueblood, C. (1986, April). Factor analysis: A tool for studying mathematics anxiety. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA. (ERIC Document Reproduction Service No. ED 270 497)

Onwuegbuzie, A. J. & Daniel, L. G. (2000, November). Reliability generalization: The importance of considering sample specificity, confidence intervals, and subgroup differences. Paper presented at the annual meeting of the Mid-South Educational Research Association, Bowling Green, KY.

*Plake, B., & Parker, C. (1982). The development and validation of a revised version of the Mathematics Anxiety Rating Scale. Educational and Psychological Measurement, 42, 551-557.

Reinhardt, B. (1996). Factors affecting coefficient alpha: A mini Monte Carlo study. In B. Thompson (Ed.), Advances in Social Science Methodology (Vol. 4, pp. 3-20). Greenwich, CT: JAI Press.

*Richardson, F., & Suinn, R. (1972). The Mathematics Anxiety Rating Scale: Psychometric data. Journal of Counseling Psychology, 19, 551-554.

*Rounds, J., & Hendel, D. (1979, April) Factor structure of the Mathematics Anxiety Rating Scale. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA. (ERIC Document Reproduction Service No. ED 180 771)

*Rounds, J., & Hendel, D. (1980). Measurement and dimensionality of mathematics anxiety. Journal of Counseling Psychology, 27,138-149.

*Satake, E., & Amato, P. (1995). Mathematics anxiety and achievement among Japanese elementary school students. Educational and Psychological Measurement, 55, 1000-1007.

Sawilowsky, S. S. (2000). Psychometrics versus datametrics: Comment on Vacha-Haase's "Reliability Generalization" method and some EPM editorial policies. Educational and Psychological Measurement, 60, 157-173.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 62, 529-540.

Suinn, R. (1988). Mathematics Anxiety Rating Scale-E (MARS-E). Fort Collins, CO: Rocky Mountain Behavioral Science Institute.

*Suinn, R., Edie, C., Nicoletti, J., & Spinelli, P. (1972). The MARS, a measurement of mathematics anxiety: Psychometric data. Journal of Clinical Psychology, 28, 373-375.

Suinn, R., & Edwards, R. (1982). The measurement of mathematics anxiety: The mathematics anxiety rating scale for adolescents: MARS-A. Journal of Clinical Psychology 38, 576-580.

*Suinn, R., Taylor, N., & Edwards, R. (1988). Suinn Mathematical Anxiety Rating Scale for elementary school students (MARS-E): Psychometric and normative data. Educational and Psycological Measurement, 48, 979-986.

*Suinn, R., Taylor, S., & Edwards, R. (1989). The Suinn Mathematics Anxiety Rating Scale (MARS-E) for Hispanic elementary school students. Hispanic Journal of Behavioral Sciences, 11, 83-90.

Thompson, B. (1994). Guidelines for authors. Educational and Psychological Measurement, 54, 837-847.

Thompson, B. (1999, February). Understanding coefficient alpha, really. Paper presented at the annual meeting of the Educational Research Exchange, College Station, TX.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. Educational and Psychological Measurement, 60, 174-195.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score

reliability across studies. Educational and Psychological Measurement, 58, 6-20.

Vacha-Haase, T., Kogan, L.R., & Thompson, B. (2000). Sample compositions and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. Educational and Psychological Measurement, 60, 509-522.

Viswesvaran, C., & Ones, D. S. (2000). Measurement error in "Big Five Factors" personality assessment: Reliability generalization across studies and measures. Educational and Psychological Measurement, 60, 224-235.

Wigfield A., & Meece, J. (1988), Math anxiety in elementary and secondary school students. Journal of Educational Psychology 80, 210-216.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, 54, 594-604. [reprint available through the APA Home Page: http://www.apa.org/journals/amp/amp548594.html]

*Wilson, V. (1997, November). Factors related to anxiety in the graduate statistics classroom. Paper presented at the annual meeting of the Mid-South Educational Research Association, Memphis, TN. (ERIC Document Reproduction Service No. ED 415 288)

Yin, P., & Fan, X. (2000). Assessing the reliability of Beck

Depression Inventory scores: Reliability generalization across studies. Educational and Psychological Measurement, 60, 201-223.

Table 1

MARS Score Reliability Estimates Across Studies

Reliability	M	SD	Min.	Max.	n
Overall	.900	.086	.550	.998	35
alpha	.915	.083	.550	.998	28
Test-retest	.841	.073	.720	.950	7

Table 2

Descriptive Statistics for Model Predictors in Regression Analyses

	Model 1		Model 2		Model 3		Model 4	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Number of Items	60.57	37.26	52.25	36.91	46.22	35.83	35.87	29.48
Likert Scale	4.83	.38	4.79	.42	4.67	.49	4.60	.51
Sample Size	394.86	399.24	433.96	415.38	314.83	375.70	321.27	381.82
Mixed Ages	.02	.17	.04	.19	.06	.24	.07	.26
Children	.20	.41	.21	.42	.28	.46	.33	.49
Adolescents	.14	.36	.18	.39	.17	.38	.20	.41
College	.51	.51	.43	.50	.33	.49	.20	.41
Adults	.11	.32	.14	.36	.17	.38	.20	.41
Reliability Type	1.20	.41	NA	NA	1.17	.38	NA	NA
Total Score *SD*	NA	NA	NA	NA	21.41	21.47	21.21	17.68
n	35		28		18		15	

Table 3

Standardized Regression, Structure Coefficients, and Effects for Regression Analyses

	Model 1		Model 2		Model 3		Model 4	
Variable/Statistic	β	r_s	β	r_s	β	r_s	β	r_s
Number of Items	.021	.137	.144	.564	-.272	-.346	.223	.354
Likert Scale	-.138	-.128	.118	-.005	.131	-.111	.659	.215
Sample Size	.080	.266	-.211	.117	-.096	-.073	-.681	-.317
Mixed Ages	-.133	.064	.041	.022	-.200	.067	.268	-.028
Children	-.330	-.159	-.046	-.051	-.504	.112	NA	NA
Adolescents	-.247	.122	-.016	.014	-.397	.018	.093	-.211
College	NA	NA	NA	NA	NA	NA	.185	.844
Adults	-.522	-.602	-.572	-.905	-.652	-.123	-.629	-.422
Reliability Type	-.488	-.545	NA	NA	-.955	-.790	NA	NA
Total Score SD	NA	NA	NA	NA	.130	-.323	-.168	.408
R^2	.404		.332		.644		.582	
n	35		28		18		15	

Note. β = standardized regression coefficient. r_s = structure coefficient.